7.7 A 56nm CMOS 99mm² 8Gb Multi-level NAND Flash Memory with 10MB/s Program Throughput

Ken Takeuchi[1], Yasushi Kameda[1], Susumu Fujimura[1], Hiroyuki Otake[1], Koji Hosono[1], Hitoshi Shiga[1], Yoshihisa Watanabe[1], Takuya Futatsuyama[1], Yoshihiko Shindo[1], Masatsugu Kojima[1], Makoto Iwai[1], Masanobu Shirakawa[1], Masayuki Ichige[1], Kazuo Hatakeyama[1], Shinichi Tanaka[1], Teruhiko Kamei[2], Jia-Yi Fu[2], Adi Cernea[3], Yan Li[3], Masaaki Higashitani[3], Gertjan Hemink[2], Shinji Sato[2], Ken Oowada[2], Shih-Chung Lee[2], Naoki Hayashida[2], Jun Wan[3], Jeffrey Lutze[3], Shouchang Tsao[3], Mehrdad Mofidi[3], Kiyofumi Sakurai[1], Naoya Tokiwa[1], Hiroko Waki[1], Yasumitsu Nozawa[1], Kazuhisa Kanazawa[1], Shigeo Ohshima[1]

[1]Toshiba, Yokohama, Japan
[2]SanDisk,Yokohama, Japan
[3]SanDisk, Sunnyvale, CA

The multi-level NAND flash memory is widely used in portable devices such as digital still cameras, USB memories, MP3 players, cell phones and portable movie players, because it doubles the memory density. As the number of pixels of image sensors, the clock frequency of CPUs, and the bandwidth of wireline and wireless communication increase, there is a strong demand for fast programming. To meet such requirements, a 10MB/s 8Gb NAND flash memory is developed. The program time is almost twice as fast as those previously reported [1, 2] and comparable to binary memories. The 98.8mm² chip with a 0.0075μm²/b memory cell is realized in a 56nm CMOS process. 8kB page programming, noise-cancellation circuits, and the dual V_{DD} line scheme realize a small die size and a fast programming. An external page copy achieves a 93ms block copy, efficiently using a 1MB block size.

Figure 7.7.1 shows the micrograph of the chip. The key features are summarized in Fig. 7.7.7. The block diagram of the chip is shown in Fig. 7.7.2. The chip contains two 4Gb memory arrays. One NAND string comprises 32 cells. A block contains 128 pages and the minimum page size and block size are 4kB and 512kB, respectively. The 10MB/s programming is achieved by extending the page size from 4kB [1, 2] to 8kB. In case of the 8kB programming, the block size increases to 1MB. Although the page size is doubled, there is no area penalty, because the WL length is also doubled. As shown in Fig. 7.7.1, row decoders, page buffers, peripheral circuits and pads are located at one side. No circuit or wiring is located at three sides of the chip. Thus, an optimized floorplan with a cell area efficiency of 70% is achieved.

To improve the program speed, 3 key circuit techniques are introduced. First, noise-cancellation circuits are introduced. As a drawback of extending the WL to accelerate the programming, the long RC delay of the WL causes a serious capacitive coupling noise between the select gate and the WL and that among WLs. Furthermore, as the WL pitch is scaled down, the noise becomes more pronounced. By using noise-cancellation circuits, the noise is eliminated.

Figure 7.7.3 shows the select gate-WL noise. During the conventional read, WLs are raised to 5.5V while BLs are precharged. Then, select gates are biased to 5.5V and the charge stored in the BLs is discharged through memory cells. In the conventional scheme, although the fast ramp-up of the metal bypassed select gate enables fast access, it causes a serious capacitive coupling noise. When the select-gate rises to 5.5V, the neighboring WL rises by 1.5V, causing a read failure. In the proposed scheme, the select gate-WL noise is eliminated by raising the neighboring select gate during the BL precharge period. In case WL0 is selected, SGS rises during the BL precharge and then SGD rises after the BL precharge is completed. Although WL0 is raised by 1V due to the capacitive coupling with SGS, the bounce of WL0 diminishes during the BL precharge and no read error occurs. On the other hand, in case WL31 is selected, SGD rises during the BL precharge and WL31 rises due to the noise. Yet, the bounce of WL31 disappears during the BL precharge. When SGS rises and the BL discharge starts, the noise of selected WL31 is eliminated.

As for the inter-WL noise, during the verify read the selected WL bounces as neighboring WLs rise to 5.5V as shown in Fig. 7.7.4. It takes more than a few microseconds for the selected WL to recover, which drastically degrades the program speed. To eliminate the noise, unselected WLs are fixed at 5.5V. Consequently, the program speed is improved by 21%.

Second, an external page-copy is introduced to improve the program speed at a system level. In case of the 8kB page programming, the block size increases to 1MB. Since the block size becomes large, a fast block copy, copying data from an old block to a new block, is required to improve the system performance. This is because when a small amount of data is rewritten, a block copy frequently happens. To accelerate the block copy, a page copy is implemented as shown in Fig. 7.7.5. To copy data from the old block to the new block, data is first read from the old block. Next, because data transferred from memory cells can contain errors, data is sent off-chip for error correction. If error is detected through ECC, the corrected data comes to the memory chip as input. Then, data is programmed to the new block. By pipelining the data output and the programming, the data output and the programming are performed at the same time and a fast block copy is realized. In most cases no error is detected and the block-copy time can be accelerated to,
(Read access time+Program time) × (number of pages per block) =93ms.

Third, a dual V_{DD} line scheme is introduced. As the BL pitch decreases, the inter-BL capacitance increases which increases the total BL capacitance. Moreover, as the page size is doubled, activated BLs are doubled. As a result, a huge BL capacitance, more than 10nF, must be precharged to the internal supply voltage, V_{DD}, during the program time. In the conventional scheme, single V_{DD} line supplies all circuits. To avoid a huge V_{DD} drop that causes a malfunction of circuits, BLs are precharged slowly, which drastically degrades the program speed. In the proposed dual V_{DD} line scheme, two internal supply-voltage lines are shorted at the V_{DD} generator circuit as shown in Fig. 7.7.6. V_{DD_BL} for the BL precharge is isolated from V_{DD} that supplies other circuits. Although V_{DD_BL} drops by 0.7V during the BL precharge, the V_{DD} drop is less than 0.2V. Since the BL precharge circuit is a simple transfer gate, the huge V_{DD_BL} drop does not affect the operation. As a result, both a reliable and a fast programming are realized. By using the dual V_{DD} line scheme, the program speed improves by 16%.

Acknowledgements:
The authors appreciate H.Nakai, M.Momodomi, K.Imamiya, H.Nakamura, F.Arai, K.Shuto, H.Hazama, H.Meguro, Dan Guterman, Jian Chen, Tuan Pham, Henry Chien, Jim Chan, Farookh Moogat, Yupin Fong, and the entire design, layout, CAD, device, and process team.

References:
[1] T. Hara, et al., "A 146mm² 8Gb NAND Flash Memory with 70nm CMOS Technology," *ISSCC Dig. Tech..Papers*, pp. 44-45, Feb., 2005.
[2] D.S. Byeon, et al., "A 8Gb Multi-Level NAND Flash Memory with 63nm STI CMOS Process Technology," *ISSCC Dig. Tech. Papers*, pp. 46-47, Feb., 2005.
[3] J. D. Lee, et al., "Effects of Parasitic Capacitance on NAND Flash Memory Cell Operation," *NVSMW Tech. Dig.*, pp. 90-91, Aug., 2001.

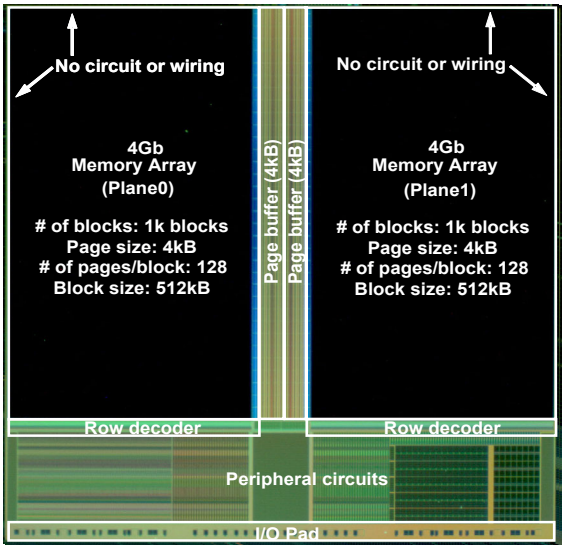

Figure 7.7.1: Die micrograph.

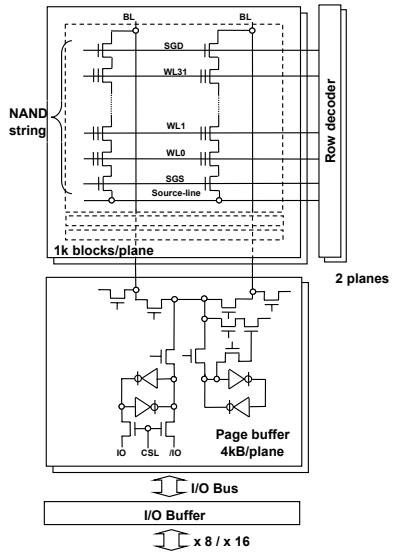

Figure 7.7.2: Block and schematic diagram of the chip.

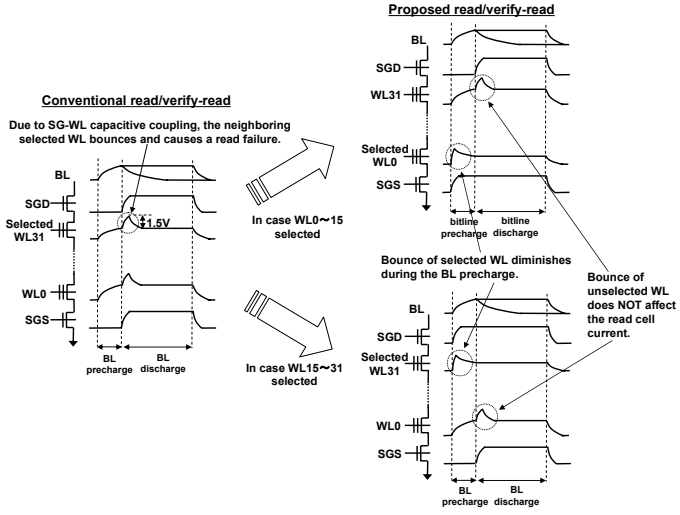

Figure 7.7.3: Select gate-WL noise-cancellation circuits.

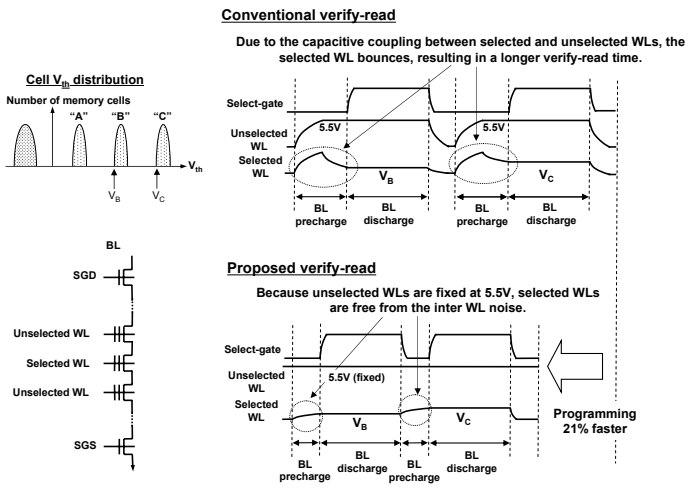

Figure 7.7.4: Inter-WL noise-cancellation circuits.

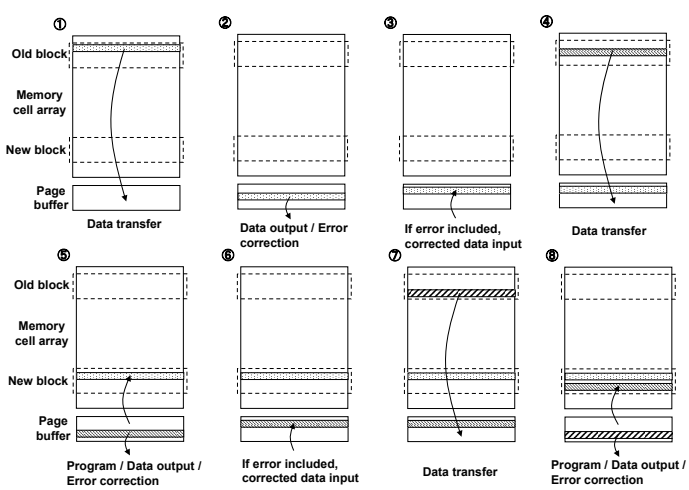

Figure 7.7.5: External page-copy function.

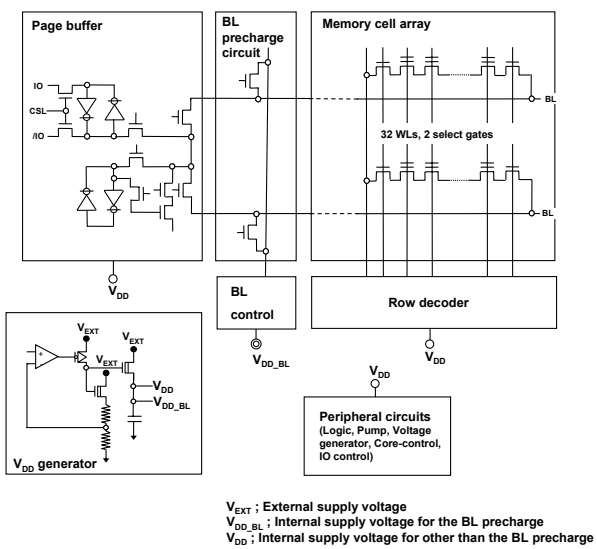

Figure 7.7.6: Dual-V_{DD} line scheme.

Continued on Page 645

Technology	3M 56nm CMOS
Cell size	0.0075μm² / b (effective)
Chip size	98.8mm²
Organization	4314 x 128pages x 1kblocks x 8 2157 x 128pages x 1kblocks x 16
Power supply	2.7～3.6V
Read access time	50μs transfer (maximum)
Burst cycle time	30ns (EDO； 50pF)
Program time	680μs (typical)
Erase time	2ms (typical)

Figure 7.7.7: Key features of the chip.

